09 - Multivariate Two Sample Inference

Junvie Pailden

SIUE, F2017, Stat 589

September 19, 2017

Comparing Mean Vectors from Two Populations

Sample	Mean	Covariance
1	$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{x}_{1j}$	$\mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)'$
2	$\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{x}_{2j}$	$\mathbf{S}_1 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'$

Let $\mu_1 = E(\mathbf{X}_1)$ and $\mu_2 = E(\mathbf{X}_2)$.

We want to answer the questions

1. Is $\mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$)?
2. If $\mu_1 \neq \mu_2$, which component means are different?

Assumptions on the Structure of the Data

- The sample $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ is a random sample from a population with mean μ_1 and covariance matrix Σ_1.
- The sample $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$ is a random sample from a population with mean μ_2 and covariance matrix Σ_2.
- The sample $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ are independent from $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$.
- For small sample sizes, the populations are multivariate normal.
- Suppose $\Sigma_1 = \Sigma_2 = \Sigma$.

- We can pool the information in both samples in order to estimate the common variance Σ.

$$\mathbf{S}_{\text{pooled}} = \frac{\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)' + \sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'}{n_1 + n_2 - 2}$$
$$= \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2$$

Test the hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$

- To test H_0, we consider the squared statistical distance from $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ to δ_0. \item By the independence assumption,

$$Cov(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = Cov(\bar{\mathbf{X}}_1) + Cov(\bar{\mathbf{X}}_2) = \frac{1}{n_1}\Sigma + \frac{1}{n_1}\Sigma$$

- Because $\mathbf{S}_{\text{pooled}}$ estimates Σ, then

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{\text{pooled}}$$

is an estimator of $Cov(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$.

Test the hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$

- The likelihood ratio test of $H_0 : \mu_1 - \mu_2 = \delta_0$ is based on the the square of the statistical distance, T^2. Reject H_0 if

$$
\begin{aligned}
T^2 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{\text{pooled}} \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0) \\
&> \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha) = c^2
\end{aligned}
$$

Test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$

If $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ is a random sample of size n_1 from $N_p(\mu_1, \Sigma)$ and $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$ is an independent random sample of size of n_2 from $N_p(\mu_2, \Sigma)$, then

$$T^2 = [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{\text{pooled}} \right]^{-1} [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2)]$$

is distributed as

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}.$$

Consequently,

$$P \left[T^2 \leq \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha) \right] = 1 - \alpha$$

Simultaneous Confidence Intervals

With probability $1-\alpha$,

$$\mathbf{a}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \pm c\sqrt{\mathbf{a}'\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{S}_{\textsf{pooled}}\mathbf{a}}$$

will cover $\mathbf{a}'(\mu_1 - \mu_2)$ for all $\mathbf{a}'$. In particular $\mu_{1i} - \mu_{2i}$ will be covered by

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm c\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{ii,pooled}} \quad \text{for } i = 1, 2, \ldots, p$$

Example: Bird Data

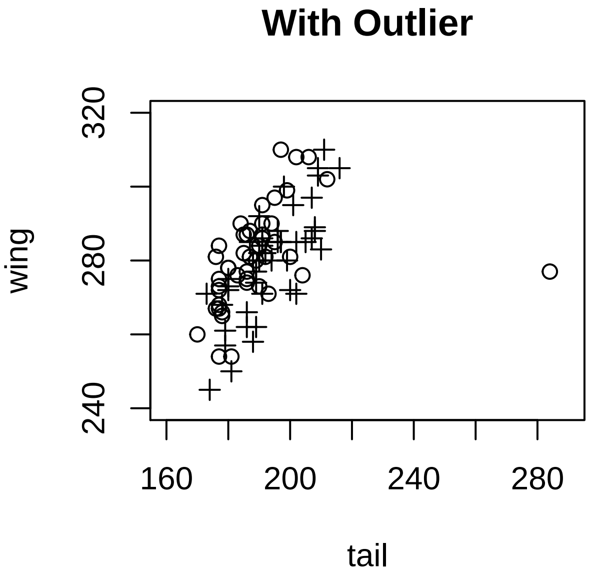
The tail lengths in millimeters (x_1) and wing lengths in millimeters (x_2) for 45 male hook-billed kites are given in file **T6-11.DAT**. Similar measurements for female hook-billed kites were given **T5-11.DAT**.

- Plot the male hook-billed kite data as a scatter diagram, and (visually) check for outliers.
- Test for equality of mean vectors for the populations of male and female hook-billed kites. Set $\alpha - .05$. If $H_0 : \mu_1 - \mu_2 = 0$ is rejected, find the linear combination most responsible for the rejection of H_0.
- Determine the 95% confidence region for $\mu_1 - \mu_2$ and 95% simultaneous confidence intervals for the components of $\mu_1 - \mu_2$.
- Are male or female birds generally larger?

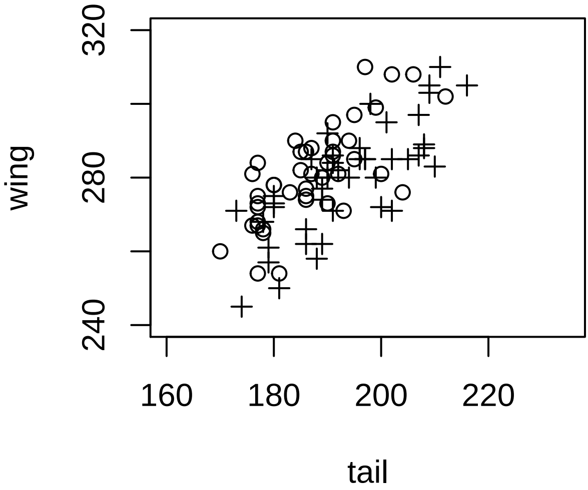
```r
bird.females <- read.table("T5-12.DAT", header = F)
bird.males <- read.table("T6-11.DAT", header = F)
colnames(bird.females) =
    colnames(bird.males) =  c("tail","wing")
```

```r
# Plot data:
plot(bird.males, main = "With Outlier",
     xlim = c(160,290), ylim=c(240,320)
)
points(bird.females, pch=3)

# Remove outlier for males and plot data:
newbird.males <- bird.males[-31, ]
plot(newbird.males, main = "Without Outlier",
        xlim=c(160,235), ylim=c(240,320))
points(bird.females, pch=3)
```


With Outlier
wing
tail
320
280
240
160
200
240
280

Without Outlier



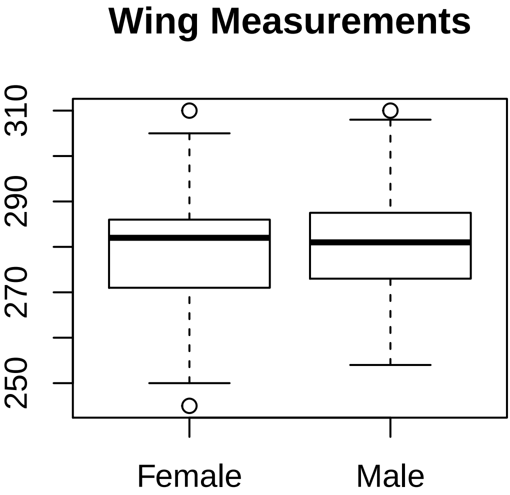
Boxplots

```
n1 <- nrow(newbird.males); n2 <- nrow(bird.females)
gender <- c(rep("Male", n1), rep("Female", n2))
new.dat <- cbind(gender,
                 rbind(newbird.males, bird.females))
str(new.dat)
```

```
# 'data.frame': 89 obs. of  3 variables:
#  $ gender: Factor w/ 2 levels "Female","Male": 2 2 2 2 2
#  $ tail  : int  180 186 206 184 177 177 176 200 191 193
#  $ wing  : int  278 277 308 290 273 284 267 281 287 271
```

```
boxplot(tail ~ gender, data = new.dat, main = "Tail Measure
```

```r
boxplot(wing ~ gender, data = new.dat, main = "Wing Measure
```



Multivariate Energy Normality Tests

```r
energy::mvnorm.etest(newbird.males, R = 199) # male
```

```
#
#   Energy test of multivariate normality: estimated parame
#
# data:  x, sample size 44, dimension 2, replicates 199
# E-statistic = 0.7, p-value = 0.3
```

```r
energy::mvnorm.etest(bird.females, R = 199) # female
```

```
#
#   Energy test of multivariate normality: estimated parame
#
# data:  x, sample size 45, dimension 2, replicates 199
# E-statistic = 0.6, p-value = 0.5
```

Summary Measures, Need `mosaic` Package

```
mosaic::favstats(tail ~ gender, data = new.dat)
```

```
#   gender min  Q1 median  Q3 max mean   sd  n missing
# 1 Female 173 187    191 202 216  194 11.0 45       0
# 2   Male 170 178    186 192 212  187  9.4 44       0
```

```
mosaic::favstats(wing ~ gender, data = new.dat)
```

```
#   gender min  Q1 median  Q3 max mean sd  n missing
# 1 Female 245 271    282 286 310  280 14 45       0
# 2   Male 254 273    281 287 310  281 13 44       0
```

Multivariate Two-Sample Tests

$H_0 : \mu_1 - \mu_2 = 0$ (no difference between means)

```
ICSNP::HotellingsT2(X = newbird.males, Y = bird.females)
```

```
#
#   Hotelling's two sample T2-test
#
# data:  newbird.males and bird.females
# T.2 = 10, df1 = 2, df2 = 90, p-value = 2e-05
# alternative hypothesis: true location difference is not e
```

Reject H_0 at 1% level of significance. Strong evidence in the sample supporting the claim that the mean measurements between genders are different.